ST494/ST694 Statistical Learning
Assignment 3

Due: 11:59pm on Friday, February 14, 2025

ST494: Questions 8-9
ST694: Questions 8-9

Part I: Practice Questions note: the page number and question number follow the newest version of the textbook.

1. Page 283, Problem 2, "For parts (a) through (c) indicate which of i through iv. is correct. Justify your answer."
2. Page 285, Problem 5, "It is well known that ridge regression tends to give...."
3. Page 287, Problem 9, "In this exercise, we will predict the number of applications received using,..."
4. Page 321, Problem 1, "It was mentioned in the chapter that a cubic regression spline with one knot...."
5. Page 323, Problem 4, "Suppose we fit a curve with basis functions $b_1(X) = I(0 \leq X \leq 2) - (X-1)I(1 \leq X \leq 2)$..."
6. Page 324, Problem 9, "This question uses the variables **dis** (the weighted mean of distances to five Boston employment centres) and ....."
7. Page 324, Problem 10, "This question relates to the *College* data set...."

Part II: Hand-in Questions

8. We wish to predict a baseball player's salary on the basis of various statistics associated with performance in the previous year. The data set "Hitters" is from *library(ISLR)*. We use function *na.omit(Hitter)* to remove all of the rows that have missing values in any variable. Please use "Salary" as the response variable and all the continuous variables as explanatory variables to implement variable selection algorithms and to find the best model for this data. You may need to use the below R codes.

```
library(ISLR)
data(Hitters)

Hitters.new <- na.omit(Hitters)

X <- as.matrix(Hitters.new[,c(-14,-15,-20,-19)])
Y <- as.matrix(Hitters.new[,19])
```

(1). [1 mark] Use C_p method to find the best model where p is the total number of parameters.

(2). [2 marks] Use the best subset selection method with two different variable selection criteria: R^2_{adj} and BIC to choose the best models.

(3). [4 marks] Use *step()* function to implement "backward elimination", "forward selection" and "hybrid" methods to choose the best models.

9. In this question you will use the famous "iris" data, which consists of four measurements on each of 150 iris flowers. They are the sepal length and width, and the petal length and width. There are three species of flowers (setosa, versicolor, and virginica), and 50 of each species have been measured.

(a). By calculating the covariance matrix of the four variables (without any scaling) and decomposing this into eigenvalues and eigenvectors, show the principal component directions and the corresponding variances of the data when projected onto those directions. Useful R functions include **eigen**, **var**, which calculate an eigen-decomposition and a covariance matrix. Compare these to the results from the built-in function **princomp** in the "stats" library.

(b). Interpret the first two principal components of the data. That is, look at the four coefficients that make up the first direction, and comment on an interpretation of this linear combination, and do the same for the second set of four coefficients.

(c). Would it be appropriate to scale the four variables before doing a PCA? Provide one argument for scaling and one against.

(d). Plot the data projected onto the first two principal components. Does the pattern you notice have anything to do with the three species of flower?

(e). Calculate a separate principal components analysis for each of the three species. Comment on how the first principal component varies across species, and provide a practical interpretation of the first principal component in each of the three groups. Suggest a reason for why the results might change when the principal components are calculated within each class separately.